

Semi-supervised learning of local structured output predictors

Xin Du^a

^a*School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China*

Abstract

In this paper, we study the problem of semi-supervised structured output prediction, which aims to learn predictors for structured outputs, such as sequences, tree nodes, vectors, etc., from a set of data points of both input-output pairs and single inputs without outputs. The traditional methods to solve this problem usually learn one single predictor for all the data points, and ignore the variety of the different data points. Different parts of the data set may have different local distributions, and require different optimal local predictors. To overcome this disadvantage of existing methods, we propose to learn different local predictors for neighborhoods of different data points, and the missing structured outputs simultaneously. In the neighborhood of each data point, we proposed to learn a linear predictor by minimizing both the complexity of the predictor and the upper bound of the structured prediction loss. The minimization is conducted by gradient descent algorithms. Experiments over four benchmark data sets, including DDSM mammography medical images, SUN natural image data set, Cora research paper data set, and Spanish news wire article sentence data set, show the advantages of the proposed method.

Keywords: Machine learning, Structured output, Semi-supervised learning, Local linear regression, Gradient descent

1. Introduction

Machine learning refers to the problem of learning a predictive model to predict a output from a input data point [1, 2, 3, 4, 5, 6, 7, 8, 9, 10,

Email address: xindu.njtu@gmail.com (Xin Du)

11, 12]. The forms of output are various, usually including binary class label and continues response. The problem of predicting binary class label is called classification [13, 14, 15, 16, 17, 18, 19, 20, 21], while the problem of predicting continues response is called regression [22, 23, 24, 25]. Both of these two problems have many applications, such as computer vision, natural language processing, bioinformatics, and finance. However, in many these applications, the forms of outputs of the prediction may be beyond binary class labels and continues responses. For example, in the part-of-speech tagging problem of natural language processing, given a sequence of words, we want to predict the tags of the part-of-speech of the words, and the output of the prediction is a sequence of parts-of-speech [26, 27, 28, 29, 30]. In the problem of hierarchical image classification problem, the class labels of images are organized as a tree structure, and the outputs of the prediction problem are the leaves of a tree [31, 32, 33, 34]. In this case, the predictive models designed for binary class labels and continues responses cannot handle these output forms, and new predictive model should be developed. The output forms other than binary labels and continues responses are called structured outputs. The structured outputs include a wide range of types of outputs, such as sequences, vectors, graph nodes, tree leaves, etc. The problem of learning predictive models to predict unknown structured outputs are called as structured output prediction [35, 36, 37, 38, 39]. Given a set of input-output data points, where the outputs are structured, this problem usually learns a predictive model to match the input-output relationship. Most existing methods designed to solve assumes that in the training set, all the input data points have their corresponding outputs. However, in real-world application, many output are not available for the inputs [40, 41, 42, 43]. The training set is composed of two parts. One part is a set of input-output pairs, which is called labeled set. The other part is a set of single input data points with missing outputs, and this set is called unlabeled data set. Learning from such a training set is call semi-supervised learning [44, 45, 46]. In this paper, we invest the problem of learning structured output predictors from such a training set. This problem is called semi-supervise structured output prediction.

1.1. Related works

Our work is a novel semi-supervised structured output prediction method, thus we introduce the related works of this direction. There are a number of existing algorithms for this problem, which are briefly introduced as follows.

1. Altun et al. [40] proposed the problem of semi-supervised learning with structured outputs. Moreover, a novel discriminative approach was also proposed to use the manifold of input features of both labeled and unlabeled data points. This approach is based on the semi-supervised maximum-margin formulation. It is an inductive algorithm, and it can be easily extended to new-coming test data points.
2. Brefeld and Scheffer [41] proposed to solve the problem of semi-supervised structured output prediction by learning in the space of input-output space, and using co-training method. This method is based on the assumption that the multiple structured output predictors should be consent with each other. Based on this assumption, the structural support vector machine is extended to the argued input-output space.
3. Suzuki et al. [42] proposed a hybrid method to solve the problem of semi-supervised structured output learning. This method combines both the generative and discriminative methods. The objective of this method is composed of log-linear forms of both discriminative structured predictor and generative model. The generative model is used to incorporate unlabeled data points. The discriminant functions is enhanced by the unlabeled data points provided by the generative model.
4. Jiang et al. [43] proposed to regularize the structured outputs by the manifold constructed from the input space directly. This method constructs a nearest neighbor graph from the input features, and use it to represent the manifold. Then the manifold is used to regularize the learning of the missing outputs of the unlabeled data points. The outputs and the predictor are learned simultaneously, and they regularize each other in the learning process.

Our work approximate the upper bound of the structured loss, and is inspired by the lower bound approximation of the structure learning of the Bayesian network [16, 4]. Thus we also discuss the works of bound approximation technologies of [16, 4].

1. Fan et al. [16] proposed to tighten the upper and lower bounds of the breadth-first branch and bound algorithm for the learning of Bayesian network structures. The informed variable groupings is used to create the pattern databases to tighten the lower bounds, while the anytime learning algorithm is used to tighten the upper bound. These strategies show good performance in the learning process of the Bayesian

network structures. The work of [16] is a contribution of major significance to the bound approximation community, and our upper bound approximation method is also based on these strategies.

2. Fan et al. [4] further proposed to improve the lower bound function of static k-cycle conflict heuristic for the learning of Bayesian network structures. This work is used to guide the search of the most promising search spaces. It uses a partition of the random variables of a data set, and the further research is based on the importance of the partition. A new partition method was proposed, and it uses the information extracted from the potentially optimal parent sets.

1.2. Our contributions

All the mentioned semi-supervised structured output prediction methods learn one single predictor for the entire data set. However, we observe that on a training set, the local distributions of neighborhoods play important roles in the problem of modeling of both input and structured outputs. It is extremely important to respect the local distributions when the structured output predictors are learned. This is even more important for learning from semi-supervised data sets. This is because for this type of data set, only a few data points have available structured outputs, and the structured outputs of all other data points are missing. To learn the missing structured outputs, we need to explore the connections between different data points, so that we may propagate the structured outputs from the labeled set to the unlabeled set. It has been shown that using local connections is an effective way to model the connections among different data points [43]. To explore the local distributions, one option is to construct a nearest neighbor graph, and use it to regularize the learning of the predictors. More specifically, with the nearest neighbor graph, we hope that the neighboring data points can obtain similar structured outputs from the predictor [47, 43]. However, one single predictor is usually not enough to characterize multiple local distributions, thus even we use the neighborhood graph to regularize the learning of the predictor, it is still not guaranteed that the local distributions are sufficiently modeled with regard to the structured output prediction problem. This is an even more serious problem when it is applied to a semi-supervised data set. With such a data set, only a few data points have corresponding structured outputs, while most of the data points do not. The learned predictor can easily fit to the labeled data points.

To solve this problem, we propose to learn multiple local linear structured output predictor for different neighborhoods to model the local distributions, instead of learning one single predictor for the entire data [48, 49]. Moreover, we also propose to learn the missing structured outputs for a semi-supervised data set simultaneously. For each data point, we propose to present the local distribution around this data point by its k nearest neighborhood, and model it by learning a local linear structured output predictor. To learn the parameters of this local predictor, we propose to minimize a upper bound of the structured losses of the data points in this neighborhood, and the squared ℓ_2 norm of the predictor parameter vector. In this process, the predicted structured outputs are compared to the learned structured outputs. The learning of the structured outputs are simultaneously by the true structured outputs of the labeled data points, and the local predictors. Some data points are shared by different neighborhoods, and they play the role of bridging different local distributions to learn a complete manifold. To solve the problem, we develop an iterative algorithm, by using gradient descent method.

1.3. Paper organization

The rest parts of this paper are organized as follows. In section 2, we model the learning problem, and present the optimization methods for the problem. In section 3, the iterative algorithm for the learning process, and the algorithm for the test process are both introduced. In section 4, the experiments in four benchmark data sets are presented, including DDSM Mammography medical image data set, SUN natural image data set, Cora research paper data set, and Spanish news wire article sentence data set. In section 5, the paper is concluded. In section 6, we discuss the future works.

2. Problem modeling and optimization

2.1. Problem modeling

Suppose we have a training set of n data points, $\mathcal{X} = \mathcal{L} \cup \mathcal{U}$, which is composed of a labeled subset, \mathcal{L} , and a unlabeled subset, \mathcal{U} . \mathcal{L} contains l data points of input-output pairs, $\mathcal{L} = \{(\mathbf{x}_i, \bar{y}_i)\}_{i=1}^l$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the d -th dimensional feature vector input of the i -th data point, $\bar{y}_i \in \mathcal{Y}$ is the true structured output of the i -th data point, \mathcal{Y} is the space of structured outputs. \mathcal{U} contains $u = n - l$ data points of inputs without outputs, $\mathcal{U} = \{\mathbf{x}_i\}_{i=l+1}^n$. The structured output prediction problem is to learn predictors to predict the structured outputs from the inputs from the training set. We

proposed to learn a local predictor for the neighborhood of each data points, instead of learning one single predictor for all the data points. We present the neighborhood of the i -th data point as the set of its k nearest neighbors,

$$\mathcal{N}_i = \{\mathbf{x}_j : \mathbf{x}_j \text{ is the } l\text{-th nearest neighbor of } \mathbf{x}_i, \text{ and } l \leq k\}. \quad (1)$$

Given a data point of the i -th neighborhood, $\mathbf{x}_j \in \mathcal{N}_i$, to predict its structured output, we match it against all the possible structured outputs. Given a candidate structured output, $y \in \mathcal{Y}$, and an input feature vector, \mathbf{x}_j , we use a joint representation to match them. The joint representation of y and \mathbf{x}_j is denoted as $\Phi(\mathbf{x}_j, y) \in \mathbb{R}^m$, where m is the dimension of the joint representation. We design a local linear predictor for the joint representations of the data points of the neighborhood of each data point, \mathcal{N}_i , and use it to predict the structured outputs,

$$y_j^* = \arg \max_{y \in \mathcal{Y}} \mathbf{w}_i^\top \Phi(\mathbf{x}_j, y), \quad j : \mathbf{x}_j \in \mathcal{N}_i, \quad (2)$$

where \mathbf{w}_i is the parameter of the local predictor of the i -th neighborhood, \mathcal{N}_i , and y_j^* is the predicted structured output of a data point in \mathcal{N}_i , \mathbf{x}_j . Apparently, this local predictor matches the joint representations of an input vector and each candidate output by a linear function, and returns the output candidate which gives the maximum matching response.

Because the training set is composed of labeled and unlabeled subsets, the outputs of the data points in \mathcal{U} are missing, we also proposed to learn the complete outputs simultaneously. The learned output set is denoted as $\{y_i\}_{i=1}^n$, where y_i is the outputs of the i -th data point. To guarantee the learned outputs are consistent with the given true outputs, we impose that the outputs of the data points in \mathcal{L} is equal to the true outputs,

$$y_i = \bar{y}_i, i : (\mathbf{x}_i, \bar{y}_i) \in \mathcal{L}. \quad (3)$$

We propose to learn the parameters of the local predictors, $\mathbf{w}_i|_{i=1}^n$, and the complete outputs, $y_i|_{i=1}^n$ simultaneously. To this end, we decompose the learning problem to each neighborhood. In the i -th neighborhood, we learn the local predictor parameter, \mathbf{w}_i , and the outputs of the data points in \mathcal{N}_i , $y_j|_{j:\mathbf{x}_j \in \mathcal{N}_i}$. We use a structured loss function to measure the loss of predicting the j -th structured output as y_j^* , while the corresponding output is y_j . The loss function is denoted as $\Delta(y_j, y_j^*)$. Moreover, we also propose to keep the

predictor as simple as possible, by minimizing the squared ℓ_2 norm of the local predictor, \mathbf{w}_i . By minimizing this loss function with regard to \mathbf{w}_i and y_j over the i -th neighborhood, and minimizing squared ℓ_2 norm of \mathbf{w}_i , we obtain the optimal local predictor parameter,

$$\begin{aligned} \min_{\mathbf{w}_i, y_j | j: \mathbf{x}_j \in \mathcal{N}_i} \left\{ \frac{1}{k} \sum_{j: \mathbf{x}_j \in \mathcal{N}_i} \Delta(y_j, y_j^*) + \frac{C}{2} \|\mathbf{w}_i\|_2^2 \right\}, \\ s.t. \ y_j = \bar{y}_j, j: (\mathbf{x}_j, \bar{y}_i) \in \mathcal{L}, \end{aligned} \quad (4)$$

where C is a tradeoff parameter to balance the weights of the first and second terms of the objective.

Because the structured loss function is usually complex and difficult to optimize directly, we seek its upper bound and minimize it instead. According to (2), we have,

$$\begin{aligned} \mathbf{w}_i^\top \Phi(\mathbf{x}_j, y_j^*) &\geq \mathbf{w}_i^\top \Phi(\mathbf{x}_j, y_j), \ j: \mathbf{x}_j \in \mathcal{N}_i, \\ \Rightarrow \mathbf{w}_i^\top (\Phi(\mathbf{x}_j, y_j^*) - \Phi(\mathbf{x}_j, y_j)) &\geq 0, \\ \Rightarrow \mathbf{w}_i^\top (\Phi(\mathbf{x}_j, y_j^*) - \Phi(\mathbf{x}_j, y_j)) + \Delta(y_j, y_j^*) &\geq \Delta(y_j, y_j^*), \\ \Rightarrow \max_{y'_j \in \mathcal{Y}} [\mathbf{w}_i^\top (\Phi(\mathbf{x}_j, y'_j) - \Phi(\mathbf{x}_j, y_j)) + \Delta(y_j, y'_j)] &\geq \Delta(y_j, y_j^*). \end{aligned} \quad (5)$$

Thus the upper bound of $\Delta(y_j, y_j^*)$ can be given as

$$\begin{aligned} \max_{y'_j \in \mathcal{Y}} [\mathbf{w}_i^\top (\Phi(\mathbf{x}_j, y'_j) - \Phi(\mathbf{x}_j, y_j)) + \Delta(y_j, y'_j)] \\ = [\mathbf{w}_i^\top (\Phi(\mathbf{x}_j, z_j^*) - \Phi(\mathbf{x}_j, y_j)) + \Delta(y_j, z_j^*)], \end{aligned} \quad (6)$$

where z_j^* is defined as

$$z_{i,j}^* = \arg \max_{y'_j \in \mathcal{Y}} [\mathbf{w}_i^\top (\Phi(\mathbf{x}_j, y'_j) - \Phi(\mathbf{x}_j, y_j)) + \Delta(y_j, y'_j)], \quad (7)$$

and it is an important parameter to define the upper bound of the loss function. With the upper bound, the minimization problem can be transferred to the following problem,

$$\begin{aligned}
\min_{\mathbf{w}_i, y_j | j: \mathbf{x}_j \in \mathcal{N}_i} & \left\{ \frac{1}{k} \sum_{j: \mathbf{x}_j \in \mathcal{N}_i} [\mathbf{w}_i^\top (\Phi(\mathbf{x}_j, z_{i,j}^*) - \Phi(\mathbf{x}_j, y_j)) + \Delta(y_j, z_{i,j}^*)] + \frac{C}{2} \|\mathbf{w}_i\|_2^2 \right\} \\
s.t. & y_j = \bar{y}_j, j : (\mathbf{x}_j, \bar{y}_j) \in \mathcal{L}.
\end{aligned} \tag{8}$$

To bridge the learning problems of different neighborhoods, we propose to combine them into one single problems over the entire data set,

$$\begin{aligned}
\min_{(\mathbf{w}_i, y_i) |_{i=1}^n} & \sum_{i=1}^n \left\{ \frac{1}{k} \sum_{j: \mathbf{x}_j \in \mathcal{N}_i} [\mathbf{w}_i^\top (\Phi(\mathbf{x}_j, z_{i,j}^*) - \Phi(\mathbf{x}_j, y_j)) + \Delta(y_j, z_{i,j}^*)] + \frac{C}{2} \|\mathbf{w}_i\|_2^2 \right\} \\
s.t. & y_i = \bar{y}_i, i : (\mathbf{x}_i, \bar{y}_i) \in \mathcal{L}.
\end{aligned} \tag{9}$$

The motive to combine the local learning problems to one single over all problem is to connect these local problems by the overlapping data points of different neighborhoods. For example, \mathcal{N}_i and $\mathcal{N}_{i'}$ have overlapping data points, and one of them is \mathbf{x}_j , i.e., $\mathbf{x}_j \in \mathcal{N}_i$, and $\mathbf{x}_j \in \mathcal{N}_{i'}$. Then the learning of y_j will be a common process of both the local problems of \mathcal{N}_i and $\mathcal{N}_{i'}$, and it can also regularize the learning of both \mathbf{w}_i and $\mathbf{w}_{i'}$.

2.2. Problem optimization

To solve the problem in (9), we propose to use an iterative algorithm. In this algorithm, we use an alternate optimization strategy. Each iteration has two steps. In the first step, we fix $y_i |_{i=1}^n$ and update $\mathbf{w}_i |_{i=1}^n$ by gradient descent algorithm. In the first step, we fix the updated $\mathbf{w}_i |_{i=1}^n$ and update $y_i |_{i=1}^n$ one by one.

2.2.1. Updating $\mathbf{w}_i |_{i=1}^n$

When the outputs $y_i |_{i=1}^n$ are fixing, the problem in (9) is transferred to

$$\begin{aligned}
\min_{\mathbf{w}_i |_{i=1}^n} & \sum_{i=1}^n \left\{ \frac{1}{k} \sum_{j: \mathbf{x}_j \in \mathcal{N}_i} [\mathbf{w}_i^\top (\Phi(\mathbf{x}_j, z_{i,j}^*) - \Phi(\mathbf{x}_j, y_j)) + \Delta(y_j, z_{i,j}^*)] \right. \\
& \left. + \frac{C}{2} \|\mathbf{w}_i\|_2^2 = g(\mathbf{w}_i) \right\},
\end{aligned} \tag{10}$$

where $g(\mathbf{w}_i)$ is the objective of the problem of the i -th neighborhood. To update \mathbf{w}_i , we use the gradient descent algorithm. However, in the objective, z_j^* is also a function of \mathbf{w}_i , thus it is difficult to obtain the gradient function directly. To solve this problem, instead of seeking the gradient function, we seek the sub-gradient function of $g(\mathbf{w}_i)$ with regard to \mathbf{w}_i . We first update each z_j^* according to previously updated \mathbf{w}_i , and then fix it to obtain the sub-gradient function with regard to \mathbf{w}_i ,

$$\nabla g(\mathbf{w}_i) = \frac{1}{k} \sum_{j:\mathbf{x}_j \in \mathcal{N}_i} [(\Phi(\mathbf{x}_j, z_{i,j}^*) - \Phi(\mathbf{x}_j, y_j))] + C\mathbf{w}_i. \quad (11)$$

With the sub-gradient function, the updating rule of \mathbf{w}_i is given as follows,

$$\begin{aligned} \mathbf{w}_i &\leftarrow \mathbf{w}_i - \eta \nabla g(\mathbf{w}_i) \\ &= \mathbf{w}_i - \eta \left(\frac{1}{k} \sum_{j:\mathbf{x}_j \in \mathcal{N}_i} [(\Phi(\mathbf{x}_j, z_{i,j}^*) - \Phi(\mathbf{x}_j, y_j))] + C\mathbf{w}_i \right) \\ &= (1 - \eta C)\mathbf{w}_i + \frac{\eta}{k} \sum_{j:\mathbf{x}_j \in \mathcal{N}_i} [(\Phi(\mathbf{x}_j, y_j) - \Phi(\mathbf{x}_j, z_{i,j}^*))]. \end{aligned} \quad (12)$$

We can see that the updating function of \mathbf{w}_i is a combination of a weighted \mathbf{w}_i and a function of the joint representations of the data points in \mathcal{N}_i .

2.2.2. Updating $y_i|_{i=1}^n$

When $\mathbf{w}_i|_{i=1}^n$ are fixed and only $y_i|_{i=1}^n$ are considered, we remove the terms irrelevant to $y_i|_{i=1}^n$ in problem of (9), and transfer it to the following problem,

$$\begin{aligned} \min_{y_i|_{i=1}^n} \quad & \sum_{i=1}^n \left\{ \frac{1}{k} \sum_{j:\mathbf{x}_j \in \mathcal{N}_i} [\Delta(y_j, z_{i,j}^*) - \mathbf{w}_i^\top \Phi(\mathbf{x}_j, y_j)] \right\} \\ \text{s.t. } & y_i = \bar{y}_i, \quad i : (\mathbf{x}_j, \bar{y}_i) \in \mathcal{L}. \end{aligned} \quad (13)$$

To solve this problem, we propose to solve the n outputs one by one. When one output y_i is considered, other outputs are fixed, $y_{i'}|_{i'=i}$. When only y_i is considered, the problem in (13) is reduced to

$$\begin{aligned} \min_{y_i} \quad & \sum_{i':\mathbf{x}_i \in \mathcal{N}_{i'}} \left\{ \frac{1}{k} [\Delta(y_i, z_{i',i}^*) - \mathbf{w}_{i'}^\top \Phi(\mathbf{x}_i, y_i)] \right\} \\ \text{s.t. } & y_i = \bar{y}_i, \text{ if } (\mathbf{x}_i, \bar{y}_i) \in \mathcal{L}. \end{aligned} \quad (14)$$

To solve this problem, we discuss it in two cases.

- **Case I**, $(\mathbf{x}_i, \bar{y}_i) \in \mathcal{L}$: In this case, according to the constraint of (16), we assign $y_i = \bar{y}_i$,

$$y_i = \bar{y}_i, \text{ if } (\mathbf{x}_i, \bar{y}_i) \in \mathcal{L}. \quad (15)$$

- **Case II**, $\mathbf{x}_i \in \mathcal{U}$: In this case, we obtain the follow optimal output as follows,

$$y_i = \arg \min_{y \in \mathcal{Y}} \sum_{i': \mathbf{x}_i \in \mathcal{N}_{i'}} \left\{ \frac{1}{k} [\Delta(y, z_{i',i}^*) - \mathbf{w}_{i'}^\top \Phi(\mathbf{x}_i, y)] \right\}, \text{ if } \mathbf{x}_i \in \mathcal{U}. \quad (16)$$

This is a minimization problem, and the objective function is a combination of functions over several neighborhoods which \mathbf{x}_i belongs to. Each function is composed of two terms, one of them is a loss function between a candidate and a upper bound parameter. The other term is a negative function of the local predictor.

3. Algorithms

3.1. Iterative learning algorithm

Based on the optimization problem, we develop an iterative algorithm to learn the local structured output predictor and the outputs jointly. In each iteration, we first fix the both the local predictor parameters and the outputs to update the upper bound parameters, then update the local predictor parameters by fixing the outputs and the upper bound parameter, and finally fix local predictor parameters and the upper bound parameters to update the outputs. The iterations are repeated for T times. The developed iterative algorithm is given in Algorithm 1.

- **Algorithm 1.** Iterative training algorithm of semi-supervised learning of local structured output predictor.
- **Inputs:** Training set, \mathcal{X} .
- **Inputs:** Maximum iteration number, T .

- **Initialize** $(\mathbf{w}_i, y_i)_{i=1}^n$;
- **For** $t = 1, \dots, T$
 - Update the upper bound parameters
 - For $i = 1, \dots, n$
 - * For $j : \mathbf{x}_j \in \mathcal{N}_i$

$$z_{i,j}^* \leftarrow \arg \max_{y'_j \in \mathcal{Y}} [\mathbf{w}_i^\top (\Phi(\mathbf{x}_j, y'_j) - \Phi(\mathbf{x}_j, y_j)) + \Delta(y_j, y'_j)], \quad (17)$$

- Update the local predictor parameters
- For $i = 1, \dots, n$

$$\mathbf{w}_i \leftarrow (1 - \eta C) \mathbf{w}_i + \frac{\eta}{k} \sum_{j: \mathbf{x}_j \in \mathcal{N}_i} [(\Phi(\mathbf{x}_j, y_j) - \Phi(\mathbf{x}_j, z_{i,j}^*))]. \quad (18)$$

- Update the structured outputs
- For $i = 1, \dots, n$
 - * If $(\mathbf{x}_i, \bar{y}_i) \in \mathcal{L}$

$$y_i \leftarrow \bar{y}_i; \quad (19)$$

* Else

$$y_i \leftarrow \arg \min_{y \in \mathcal{Y}} \sum_{i': \mathbf{x}_i \in \mathcal{N}_{i'}} \left\{ \frac{1}{k} [\Delta(y, z_{i',i}^*) - \mathbf{w}_{i'}^\top \Phi(\mathbf{x}_i, y)] \right\}; \quad (20)$$

- **Output:** $\mathbf{w}_i|_{i=1}^n$.

3.2. Algorithm of predicting structured output of test data point

We have discussed how to learn local predictors from a training set. In this section, we discuss how to use these local predictors for the task of predicting structured output of a new-coming test data point. Suppose the input feature vector of the new-coming data point is \mathbf{x} , to predict its output, we first find which neighborhoods it belongs to. To this end, we first find its k nearest neighbors from the training set, and denote the set of its k nearest neighbors as $\mathcal{N}_{\mathbf{x}}$. We assume \mathbf{x} are in the neighbors of the data points in $\mathcal{N}_{\mathbf{x}}$, and use their local predictors to predict the structured outputs of \mathbf{x} . Given a candidate output, y , we use these local predictors to match it to \mathbf{x} to obtain k matching scores. The average matching score is used as the final matching score to match \mathbf{x} and y . The candidate output which gives the maximum matching score is choose as the final structured output of \mathbf{x} ,

$$y^* = \arg \max_{y \in \mathcal{Y}} \frac{1}{k} \sum_{i: \mathbf{x}_i \in \mathcal{N}_{\mathbf{x}}} \mathbf{w}_i^\top \Phi(\mathbf{x}, y). \quad (21)$$

The prediction algorithm for new-coming data point is given in Algorithm 2.

- **Algorithm 2:** Predicting the structured output of a new-coming test data point.
- **Input:** \mathbf{x} and \mathcal{X} ;
- Find the k -th nearest neighbor set of \mathbf{x} from \mathcal{X} , $\mathcal{N}_{\mathbf{x}}$;
- **Initialize** maximum matching score $s^* = -\infty$
- **For** $y \in \mathcal{Y}$

$$s_y = \frac{1}{k} \sum_{i: \mathbf{x}_i \in \mathcal{N}_{\mathbf{x}}} \mathbf{w}_i^\top \Phi(\mathbf{x}, y); \quad (22)$$

– If $s_y \geq s^*$

$$s^* = s_y, y^* = y; \quad (23)$$

- **Output:** y^*

4. Experiments

In this section, we evaluate the performance of the proposed semi-supervised structured output prediction method. The experiments are conducted on three benchmark data sets. We first compare it to several state-of-the-art semi-supervised structured output prediction method, and then analyze its sensitivity to parameters, k and C , experimentally.

4.1. Benchmark data sets

We used three benchmark data sets in our experiments, which are discussed as follows.

- **Data set I - DDSM mammography image data set:** The last data set is a medical imaging data set which contains Mammography images [50]. This data set contains 2620 pairs of images and they belong to three different classes, normal, cancer, and benign. Thus this is a three class classification problem. Some example images of this data set are given in Fig. 1. To present the class of each image pair, we code it as a three-dimensional binary vector. For the i -th image pair, its structured output is a vector is $y_i = [y_{i1}, y_{i2}, y_{i3}]^T$, where

$$\begin{aligned} y_{i1} &= \begin{cases} 1, & \text{if the } i\text{-th image pair belongs to } \mathbf{normal} \text{ class,} \\ 0, & \text{otherwise.} \end{cases} \\ y_{i2} &= \begin{cases} 1, & \text{if the } i\text{-th image pair belongs to } \mathbf{cancer} \text{ class,} \\ 0, & \text{otherwise.} \end{cases} \\ y_{i3} &= \begin{cases} 1, & \text{if the } i\text{-th image pair belongs to } \mathbf{benign} \text{ class,} \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \quad (24)$$

To compare a given input and its corresponding prediction, y , and y^* , we use the 0-1 loss,

$$\Delta(y, y^*) = 1, \text{ if } y \neq y^*, \text{ and } 0 \text{ otherwise.} \quad (25)$$

To construct the feature vector of the i -th image pair, we use the bag-of-feature method. The images are split into small image patches, the image patches are quantized to a dictionary, and the quantization histogram is used as the feature vector of the image pair, \mathbf{x}_i . Moreover,

the joint representation of a input vector \mathbf{x} and a output vector y is their tensor product,

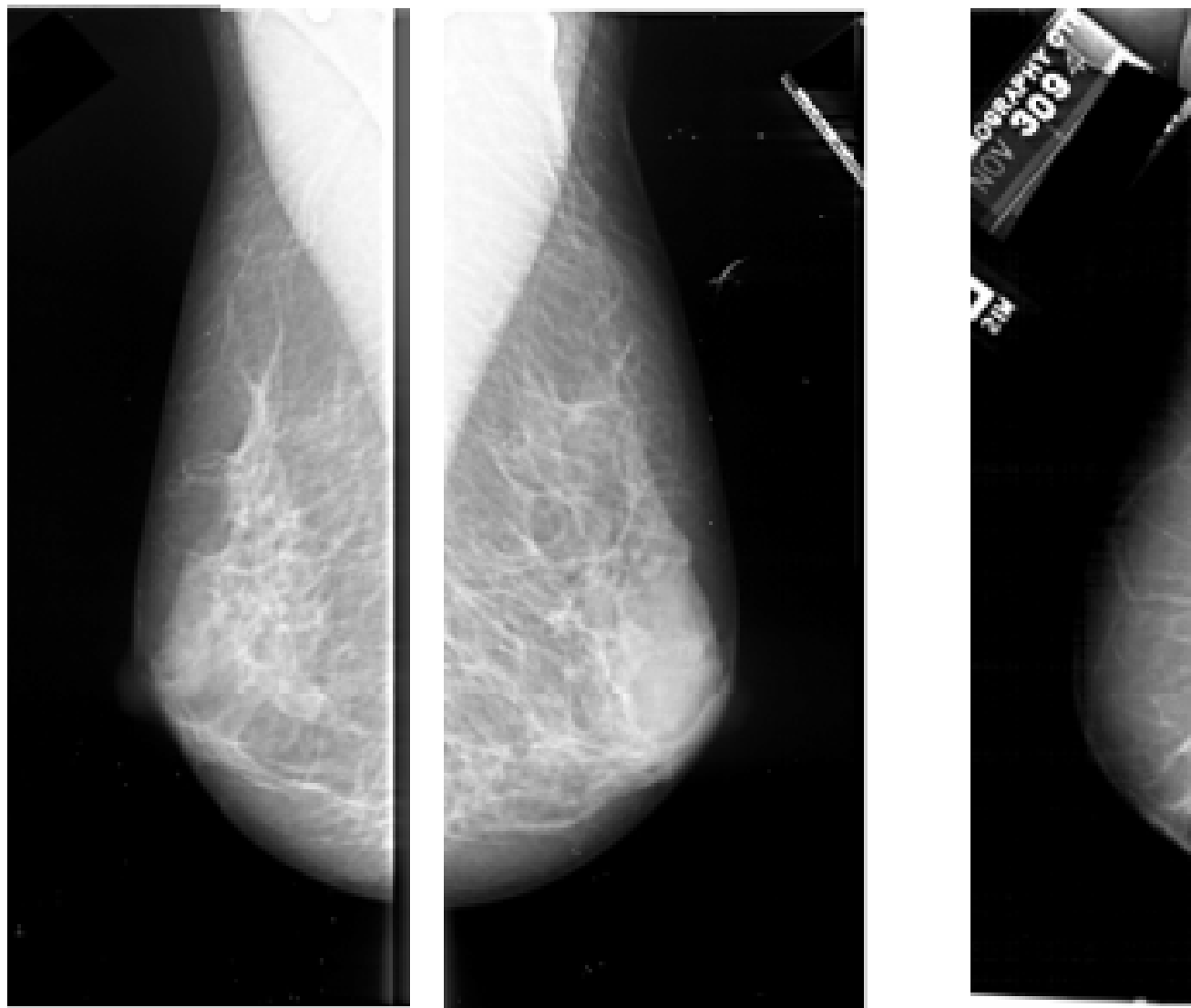
$$\Phi(\mathbf{x}, y) = \mathbf{x} \otimes y. \quad (26)$$

- **Data set II - SUN natural image data set:** The second data set is a image data set [51]. The class labels of the images of this data set is organized as a tree structure. The tree has 15 different leaves. For each class, we randomly select 200 images from the data set to conduct our experiments, thus there are 3,000 images in our data set in total. Some example images of this data set are given in Fig. 2. To extract features from each image, we calculate the HOG features. The structured output of a data point is a leave of the tree. We code it as a vector κ dimension, $y = [y_1, \dots, y_\kappa]^\top \in \{1, 0\}^\kappa$, where κ is the number of nodes of the tree. The ι -th element of the vector, $y_\iota = 1$ if the ι -th node is its predecessor or itself, and 0 otherwise. Given a input \mathbf{x} and a output y , the joint representation is

$$\Phi(\mathbf{x}, y) = \mathbf{x} \otimes y. \quad (27)$$

The loss function between y and y^* , $\Delta(y, y^*)$, is defined as the height of their first common ancestor.

- **Data set III - Cora research paper data set:** This data set is a set of computer science research papers [52]. It contains 9,947 papers. To represent the papers, we extract two types of features. The first type is the frequencies of words of the papers, and the second type is the out-links of the citations of the papers. To extract the second type features, we remover the papers which do not have links to other papers, and a data set of 9,555 is used for the experiments. These papers belong to 7 different non-overlapping classes. Thus the structured output of each paper is coded as d 7-dimensional vector. The loss function defined for this structured output is the 0-1 loss. Moreover, the joint representation of a input vector \mathbf{x} and a output vector y is also their tensor product.
- **Data set IV - Spanish news wire article sentence data set:** This data set is a data set for named entity recognition problem of natural language processing [53]. It contains 300 sentences, and each sentence



(a) normal

Figure 1: Example images of DDSM Mammography image data set.



Figure 2: Example images of SUN data set.

is used as a data point in our experiment. The length of each sentence is 9, and the corresponding output of a sentence, y , is a sequence of labels of non-name and named entities. The joint representation $\Phi(\mathbf{x}, y)$ of a sentence, \mathbf{x} , and a sequence of labels, y , is defined as the histogram of state transition, and a set of emission features. The loss function to compare y and y^* is defined as a 0-1 loss, i.e., $\Delta(y, y^*) = 1$ if $y \neq y^*$, and 0 otherwise.

4.2. Experiment setup

To conduct the experiments, we use the 10-fold cross validation strategy to split the training and test subsets. Given an entire data set, we randomly split it to ten subsets of equal sizes. Each subset is used as the test set, while the remaining nine subsets are combined and used as a training set. The training set are also randomly split to a labeled set and an unlabeled set. The Labeled set contains about 30% of the training data points, and the unlabeled set contains the remaining 70% data points. The proposed algorithm is performed to the training set to learn the local predictors, and then the local predictors are used to predict the structured outputs of the data points of the test set. The average structured loss over the test set is used to evaluate the performance of the proposed algorithm. Given a test set, \mathcal{T} with $n_{\mathcal{T}}$ test data points, the average structured loss is defined as follows,

$$Average\ Loss = \frac{1}{n_{\mathcal{T}}} \sum_{i: \mathbf{x}_i \in \mathcal{T}} \Delta(y_i, y_i^*) \quad (28)$$

where y_i is the true structured output of the i -th test data point, while y_i^* is its corresponding predicted structured output.

4.3. Comparison to state-of-the-arts

We first compare the proposed algorithm to some state-of-the-art semi-supervised structured output prediction algorithms, including the algorithms proposed by Altun et al. [40], Brefeld and Scheffer [41], Suzuki et al. [42], and Jiang et al. [43]. All the four competing algorithm learn a single global

Table 1: Average structured losses of the compared algorithms over three data sets.

Method	Data set I	Data set II	Data set III	Data set IV
SSLSOP	0.385	0.628	0.308	0.450
Jiang et al. [43]	0.401	0.677	0.383	0.492
Altun et al. [40]	0.438	0.738	0.497	0.504
Brefeld and Scheffer [41]	0.489	0.762	0.527	0.511
Suzuki et al. [42]	0.507	0.754	0.512	0.574

predictor for all the data points, and our algorithm is the only algorithm that explores the local distribution of the data set, and learns the local predictors for different neighborhoods. Our algorithm is named as semi-supervised local structured output prediction algorithm (SSLSOP). The average losses of the compared algorithms over three different data sets are given in Table 1. From the results in Table 1, it is obvious that the proposed algorithm outperforms better than all competing algorithms significantly. For example, over data set I, only SSLSOP archives an average structured loss lower than 0.400. The method of Jiang et al. [43] is the second best method. It is slightly better than the other mothers. This method also try to explore the local structure of the data set, and use it to regularize the learning of the structured output predictors. However, it still learns a global predictor, thus it cannot compete with the proposed method, SSLSOP. The results of the other algorithms are comparable to each other, and they are inferior to the SSLSOP. In summary, the results clearly show that multiple local predictors work better than one single predictor in the problems of structured output prediction.

4.4. Sensitivity to parameters

In our objective, there are two important parameters, k , the size of each neighborhood, and C , the tradeoff parameter. To study the sensitivity of the algorithm to these two parameters, we plot the curves of the results with different values of the parameters. These curves of sensitivity to parameter k over four benchmark data sets are shown in Fig. 3. From the figures, we observe that the algorithm is stable to the changes of the parameter k . For different data sets, the optimal neighborhood sizes are different. For example, over the data set I, the lowest average structured loss is obtained at $k = 20$, while for data set II, it is obtained at $k = 100$. The sensitivity curves of C are given in Fig. 4. From this figure, we also can see that the performances of the proposed algorithm are stable against the changes of the

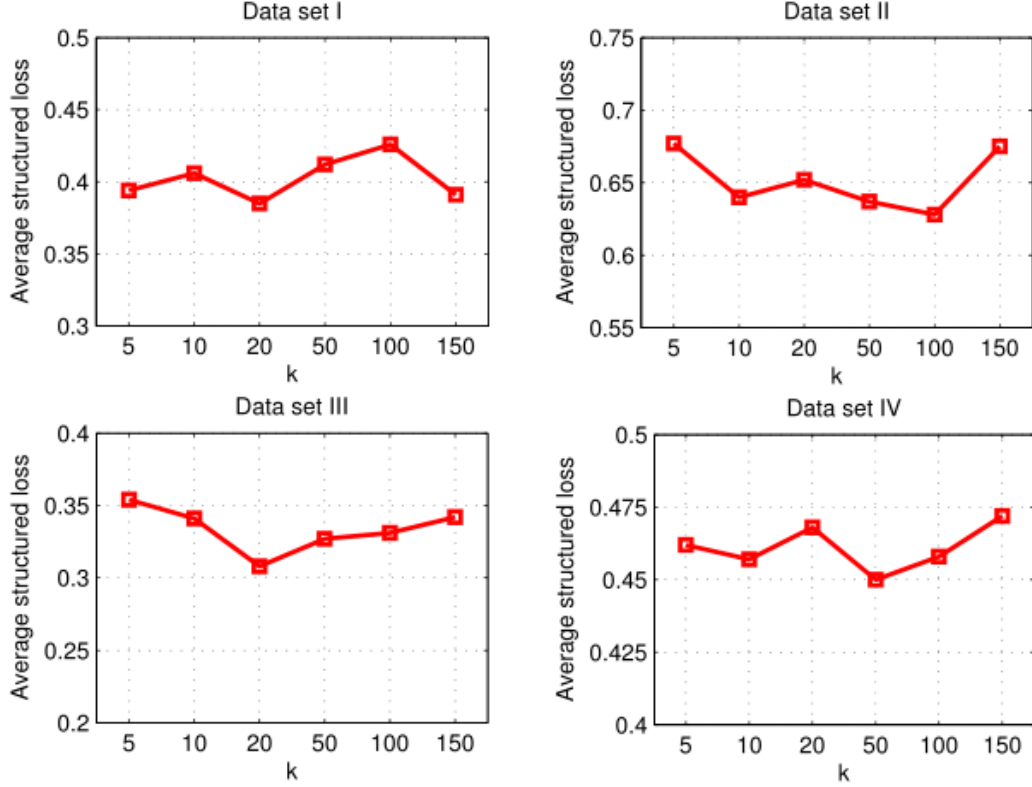


Figure 3: Curves of sensitivity of SSLSOP to parameter k .

parameter C , especially over data set II. We cannot see a clear trend of the changes of the results corresponding to the changes of the values of C . This means that the algorithm is robust to the selection of parameter C .

4.5. Repainting time analysis

We are also interested in the running time of the proposed method, SSLSOP, and its competing methods. The running time of these methods over four benchmark data sets are given in Fig. 5. From the figure, we can see that the proposed method, SSLSOP, consumes the second shortest running time over three data sets. The only exception is the results over data set II. The least time consuming algorithm is the one proposed by Altun et al. [40], however, its prediction results are not satisfying. The most time consuming algorithm is the one proposed by Jiang et al. [39], but its prediction results are not as good as SSLSOP.

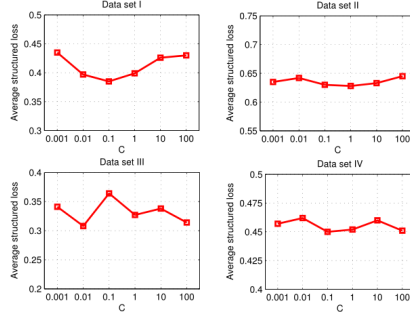


Figure 4: Curves of sensitivity of SSLSOP to parameter C .

5. Conclusion

In this paper, we investigate the problem of semi-supervised learning of structured output predictor. To handle the problem of diversity of the local distributions, we propose to learn local structured output predictors for neighborhoods of different data points. Moreover, we also propose to learn the missing outputs of the unlabeled data points. We build a new minimization problem to learn the local structured output predictors and the missing structured outputs simultaneously. This problem is modeled as the joint minimization of the local predictor complexity and the local structured output loss. The problem is optimized by gradient descent, and we designed an iterative algorithm to learn the local predictors. The experiments over benchmark data sets of medical image classification, natural image classification, computer science paper classification, and sentence part-of-speech tagging.

6. Future work

In the future, we will study how to fit the proposed algorithm to big data sets, by using big data processing framework, such as Map-Reduce of Hadoop software. When the data set is big, i.e., the number of data points, n , is large, we can use the Hadoop distributed file system (HDFS) to store the data set. The entire data set is split into some small sub-sets, and different sub-sets are stored in different clusters. The proposed algorithm has three basic steps, and each of them can be parallelized easily. The three steps are listed as follows:

1. finding the k nearest neighbors of each data point,

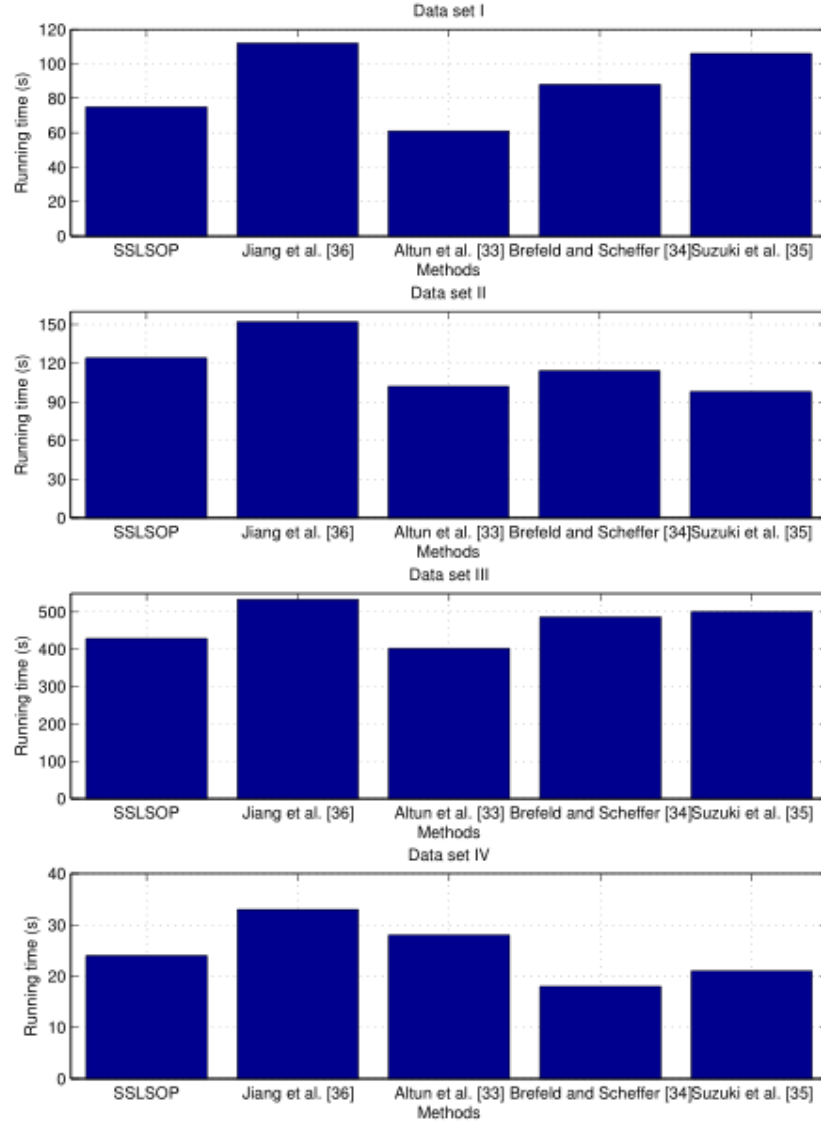


Figure 5: Running time of the compared algorithms over the four data sets.

2. updating the local structured output predictor parameter of each neighborhood, and
3. updating the outputs of each data point.

To find the k nearest neighbors of one data point from the distributed big data set, we can use the Map-Reduce framework. We use the Map function to find the k nearest neighbors from each sub-set, and then use the Reduce function to find the final k nearest neighbors from the results of the Map functions. To update the local structured output predictor parameter of one neighborhood, according to (12), only the data points of the considered neighborhood is used. We can store the data points of each neighborhood in the same cluster, and use a Map function to update the parameter simultaneously for all the local structured output predictors. Similarly, the structured outputs also are calculated by exploring the data points of the neighborhoods, by using Map functions.

References

- [1] S. Fong, S. Deb, X.-S. Yang, A heuristic optimization method inspired by wolf preying behavior, *Neural Computing and Applications* 26 (7).
- [2] G. Wang, H. Chu, Y. Zhang, H. Chen, W. Hu, Y. Li, X. Peng, Multiple parameter control for ant colony optimization applied to feature selection problem, *Neural Computing and Applications* 26 (7) (2015) 1693–1708.
- [3] X. Fan, K. Tang, T. Weise, Margin-based over-sampling method for learning from imbalanced datasets, in: *Advances in Knowledge Discovery and Data Mining*, Springer, 2011, pp. 309–320.
- [4] X. Fan, C. Yuan, An improved lower bound for bayesian network structure learning, in: *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 2439 – 2445.
- [5] S. Wang, Y. Jiang, F.-L. Chung, P. Qian, Feedforward kernel neural networks, generalized least learning machine, and its deep learning with application to image classification, *Applied Soft Computing Journal* 37 (2015) 125–141.

- [6] R. Singh, H. Kumar, R. Singla, An intrusion detection system using network traffic profiling and online sequential extreme learning machine, *Expert Systems with Applications* 42 (22) (2015) 8609–8624.
- [7] W. Li, W. Mo, X. Zhang, J. Squiers, Y. Lu, E. Sellke, W. Fan, J. Dimaio, J. Thatcher, Outlier detection and removal improves accuracy of machine learning approach to multispectral burn diagnostic imaging, *Journal of Biomedical Optics* 20 (12). doi:10.1117/1.JBO.20.12.121305.
- [8] J. Xu, W.-Q. Zhang, J. Liu, S. Xia, Regularized minimum class variance extreme learning machine for language recognition, *Eurasip Journal on Audio, Speech, and Music Processing* 2015 (1). doi:10.1186/s13636-015-0066-5.
- [9] J. Wang, H. Wang, Y. Zhou, N. McDonald, Multiple kernel multivariate performance learning using cutting plane algorithm, in: *Systems, Man and Cybernetics (SMC), 2015 IEEE International Conference on*, IEEE, 2015, pp. 1870–1875.
- [10] H. Wang, J. Wang, An effective image representation method using kernel classification, in: *2014 IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI 2014)*, 2014, pp. 853–858.
- [11] X. Liu, J. Wang, M. Yin, B. Edwards, P. Xu, Supervised learning of sparse context reconstruction coefficients for data representation and classification, *Neural Computing and Applications* (2015) 1–9.
- [12] F. Lin, J. Wang, N. Zhang, J. Xiahou, N. McDonald, Multi-kernel learning for multivariate performance measures optimization, *Neural Computing and Applications* (2016) 1–13.
- [13] F. Shu, S. Vandenberghe, J. Brackett, J. Antaki, Classification of unsteady flow patterns in a rotodynamic blood pump: Introduction of non-dimensional regime map, *Cardiovascular Engineering and Technology* 6 (3) (2015) 230–241.
- [14] X. Fan, K. Tang, Enhanced maximum auc linear classifier, in: *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*, Vol. 4, IEEE, 2010, pp. 1540–1544.

- [15] X. Fan, B. Malone, C. Yuan, Finding optimal bayesian network structures with constraints learned from data, in: Proceedings of the 30th Annual Conference on Uncertainty in Artificial Intelligence (UAI-14), 2014, pp. 200–209.
- [16] X. Fan, C. Yuan, B. Malone, Tightening bounds for bayesian network structure learning, in: Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014, pp. 2439 – 2445.
- [17] Z. Chen, W. Huang, Z. Lv, Towards a face recognition method based on uncorrelated discriminant sparse preserving projection, *Multimedia Tools and Applications*.
- [18] C. Fu, P. Zhang, J. Jiang, K. Yang, Z. Lv, A bayesian approach for sleep and wake classification based on dynamic time warping method, *Multimedia Tools and Applications* (2015) 1–20.
- [19] J. Wang, Y. Zhou, K. Duan, J. J.-Y. Wang, H. Bensmail, Supervised cross-modal factor analysis for multiple modal data classification, in: *Systems, Man and Cybernetics (SMC)*, 2015 IEEE International Conference on, IEEE, 2015, pp. 1882–1888.
- [20] J. Wang, Y. Zhou, M. Yin, S. Chen, B. Edwards, Representing data by sparse combination of contextual data points for classification, in: *Advances in Neural Networks–ISNN 2015*, Springer, 2015, pp. 373–381.
- [21] Y. Zhou, W. Hu, B. Peng, Y. Liu, Biomarker binding on an antibody-functionalized biosensor surface: The influence of surface properties, electric field, and coating density, *The Journal of Physical Chemistry C* 118 (26) (2014) 14586–14594.
- [22] S. Kavitha, S. Adish Kumar, S. Kaliappan, I. Yeom, J. Banu, Achieving profitable biological sludge disintegration through phase separation and predicting its anaerobic biodegradability by non linear regression model, *Chemical Engineering Journal* 279 (2015) 478–487.
- [23] J. Pascual-Gonzalez, C. Pozo, G. Guilln-Goslbez, L. Jimnez-Esteller, Combined use of milp and multi-linear regression to simplify lca studies, *Computers and Chemical Engineering* 82 (2015) 34–43.

- [24] H. Yang, D. Yi, Studies of the adaptive network-constrained linear regression and its application, *Computational Statistics and Data Analysis* 92 (2015) 40–52. doi:10.1016/j.csda.2015.06.008.
- [25] M. Pia-Monarez, J. Ortiz-Yaez, Weibull and lognormal taguchi analysis using multiple linear regression, *Reliability Engineering and System Safety* 144 (2015) 244–253. doi:10.1016/j.ress.2015.08.004.
- [26] V. Vincze, The relationship of dependency relations and parts of speech in hungarian, *Journal of Quantitative Linguistics* 22 (1) (2015) 44–54.
- [27] E. Fonseca, J. G Rosa, S. Alusio, Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese, *Journal of the Brazilian Computer Society* 21 (1) (2015) 14.
- [28] T. Bhowmik, N. Niu, J. Savolainen, A. Mahmoud, Leveraging topic modeling and part-of-speech tagging to support combinational creativity in requirements engineering, *Requirements Engineering* (2015) 28.
- [29] H. Carneiro, F. Frana, P. Lima, Multilingual part-of-speech tagging with weightless neural networks, *Neural Networks* 66 (2015) 11–21.
- [30] G. Wang, Z. Zhang, J. Sun, S. Yang, C. Larson, Pos-rs: A random subspace method for sentiment classification based on part-of-speech analysis, *Information Processing and Management* 51 (4) (2015) 458–479.
- [31] R. Baravalle, C. Delrieux, J. Gmez, Multifractal characterisation and classification of bread crumb digital images, *Eurasip Journal on Image and Video Processing* 2015 (1) (2015) 10.
- [32] J. Gao, L. Xu, J. Shen, F. Huang, F. Xu, A novel information transferring approach for the classification of remote sensing images, *Eurasip Journal on Advances in Signal Processing* 2015 (1) (2015) 38. doi:10.1186/s13634-015-0223-0.
- [33] B. Decost, E. Holm, A computer vision approach for automated analysis and classification of microstructural image data, *Computational Materials Science* 110 (2015) 126–133.

- [34] G. Gao, J. Yang, S. Wu, X. Jing, D. Yue, Bayesian sample steered discriminative regression for biometric image classification, *Applied Soft Computing Journal* 37 (2015) 48–59.
- [35] V. Srikumar, C. Manning, Learning distributed representations for structured output prediction, Vol. 4, 2014, pp. 3266–3274.
- [36] Y. Jiang, W. Clark, I. Friedberg, P. Radivojac, The impact of incomplete knowledge on the evaluation of protein function prediction: A structured-output learning perspective, *Bioinformatics* 30 (17) (2014) i609–i616.
- [37] Y. Han, X. Wei, X. Cao, Y. Yang, X. Zhou, Augmenting image descriptions using structured prediction output, *IEEE Transactions on Multimedia* 16 (6) (2014) 1665–1676.
- [38] N. Nourani-Vatani, R. Lpez-Sastre, S. Williams, Structured output prediction with hierarchical loss functions for seafloor imagery taxonomic categorization, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9117 (2015) 173–183.
- [39] F. Jiang, L. Jia, X. Sheng, R. LeMieux, Manifold regularization in structured output space for semi-supervised structured output prediction, *Neural Computing and Applications* (2015) 10.
- [40] Y. Altun, M. Belkin, D. A. Mcallester, Maximum margin semi-supervised learning for structured variables, in: *Advances in neural information processing systems*, 2005, pp. 33–40.
- [41] U. Brefeld, T. Scheffer, Semi-supervised learning for structured output variables, in: *Proceedings of the 23rd international conference on Machine learning*, ACM, 2006, pp. 145–152.
- [42] J. Suzuki, A. Fujino, H. Isozaki, Semi-supervised structured output learning based on a hybrid generative and discriminative approach., in: *EMNLP-CoNLL*, 2007, pp. 791–800.
- [43] F. Jiang, L. Jia, X. Sheng, R. LeMieux, Manifold regularization in structured output space for semi-supervised structured output prediction, *Neural Computing and Applications* (2015) 1–10.

- [44] H. Gan, R. Huang, Z. Luo, Y. Fan, F. Gao, Towards a probabilistic semi-supervised kernel minimum squared error algorithm, *Neurocomputing* 171 (2016) 149–155.
- [45] Z. Zeng, X. Wang, J. Zhang, Q. Wu, Semi-supervised feature selection based on local discriminative information, *Neurocomputing* 173 (2016) 102–109.
- [46] Q. Wang, K. Zhang, Z. Chen, D. Wang, G. Jiang, I. Marsic, Enhancing semi-supervised learning through label-aware base kernels, *Neurocomputing* 171 (2016) 1335–1343.
- [47] C. Hu, J. T. Kwok, Manifold regularization for structured outputs via the joint kernel, in: *Neural Networks (IJCNN), The 2010 International Joint Conference on*, IEEE, 2010, pp. 1–8.
- [48] H. Zhang, J. Yu, M. Wang, Y. Liu, Semi-supervised distance metric learning based on local linear regression for data clustering, *Neurocomputing* 93 (2012) 100–105.
- [49] H. Xue, Y. Zhu, S. Chen, Local ridge regression for face recognition, *Neurocomputing* 72 (4) (2009) 1342–1346.
- [50] M. Heath, K. Bowyer, D. Kopans, R. Moore, W. P. Kegelmeyer, The digital database for screening mammography, in: *Proceedings of the 5th international workshop on digital mammography*, Citeseer, 2000, pp. 212–218.
- [51] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba, et al., Sun database: Large-scale scene recognition from abbey to zoo, in: *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, IEEE, 2010, pp. 3485–3492.
- [52] A. McCallum, K. Nigam, L. H. Ungar, Efficient clustering of high-dimensional data sets with application to reference matching, in: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2000, pp. 169–178.
- [53] T. K. Sang, Introduction to the conll-2002 shared task: Language-independent named entity recognition, in: *Proceedings of the 6th conference on natural language learning*, pp. 155–158.